Technical Reports on Language Technology
Report No 37, 20019
http://www.njas.helsinki.fi/salama

Morphological analyser of Finnish as a finite-state transducer without rules

Arvi Hurskainen
Department of World Cultures, Box 59
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

Finnish is considered one of the most complex languages with a particularly rich morphology. In addition to complex morhology, it has also a gradation system and front/back vowel concordance. In earlier times, all these features were considered an ovewhelming obstacle for computational description of this language. The problems were succesfully resolved in 1983, when Kimmo Koskenniemi described Finnish using finite-state technology and two-level rules. In this report I describe a partly similar implemention, which, however, has an impotant difference. The system does not use morphophonological rules at all. As a result, it is a finite-state transducer.

Key Words: *morphology, finite-state technology.*

1 Introduction

Finite-state technology is a standard methodology for describing morphologically complex languages, such as Finnish. The finite-state technology, coupled with morphophonological rules ensures efficient and and fast analysis of text. The use of morphophonological rules makes it possible to simplify the lexicon structure, because variations can be handled with rules.

I had earlier implemented the analyzer of Swahili using a finite-state lexicon and two-level rules. I could have made the same for Finnish, but out of curiosity I wanted to test the possibility of describing Finnish with the finite-state lexicon only. It was clear that the lexicon would be more complex with a large number of sub-lexicons.

Much of the lexical material was already classified. The Center of National Languages (KOTUS) has classified a large number of Finnish words providing them with inflection codes. This source was a good start, and many more words were added to the list when I earlier worked with the English to Finniash translation system. I also had provided the words with front/back concordance information, which was missing in the KOTUS list.

2 Lexicon structure

In finite-state technology, words are described as a combination of morphemes. For each word, the analysis starts from the beginning of the word trying to find matching strings through the analysis system. If matches are found and the process comes to the point, where termination is allowed, the string is accepted as a valid word of the language. In the opposite case no output is given.

If no rule component is included in the system, we only need to care about the lexicon.

Because Finnish inflects with suffixes, and not with prefixes, the lexicon structure is quite straightforward. The stem of the word is the first part of the word-form, and inflectional suffixes come thereafter.

When there are no rules, it is important to branch the lexicon out to two sections accoding to the vowel concordance. Words with front vowel inflection must be kept separate from the words with back vowel inflection. As a result, the system has a total of 585 sub-lexicons.

When the number of lexical entries was over 80,000, there was a danger that the lexicon would be too large to run. Therefore, in the first phase I handled the verbs in a separate lexicon file. The lexicons were run after each other, and in between the result of the first lexicon had to be modified so that it was possible to run the second lexicon. The arrangement worked nicely.

However, in the long run it became cumbersome to handle matarials in two separate files. Then I joined the two files as a single lexicon file and tested it. After enlarging the string matrix sufficiently, the system worked fine as a single file.

3 Using underspecification for contracting the lexicon file

Because the system has hundreds of sub-lexicons for inflecting various inflection patterns, and many of them have a similar structure, there is motivation to abbreviate the description of each sub-lexicon. For example, the form may have more than one interpreteation (1).

(1)
```
"<kirjani>"
      "kirja" N SG/PL NOM/GEN/ACC-G/ACC-N POS-SG1
```

This is equal to the following format (2):

(2)
```
"<kirjani>"
      "kirja" N SG NOM POS-SG1
      "kirja" N SG GEN POS-SG1
      "kirja" N SG ACC-G POS-SG1
      "kirja" N SG ACC-N POS-SG1
      "kirja" N PL NOM POS-SG1
      "kirja" N PL GEN POS-SG1
      "kirja" N PL ACC-G POS-SG1
      "kirja" N PL ACC-N POS-SG1
```

Also, with some verb forms we can use the same method (3).

(3)
```
"<juossut>"
      "juosta" V NEG-PAST SG123
      "juosta" V PART-PERF SG123
```

```
"<juosseet>"
      "juosta" V NEG-PAST PL123
      "juosta" V PART-PERF PL123
```

Both verb forms have six interpretations each. Here they are written on two lines, although also here we could use one-line format, such as in (4).

(4)
```
"<juossut>"
      "juosta" V NEG-PAST/PART-PERF SG123
"<juosseet>"
      "juosta" V NEG-PAST/PART-PERF PL123
```

The method of using underspecification in analysis is particularly tempting in cases, where the form may belong to two POS categories (5).

(5)
```
"<juopunut>"
      "juopunut" A SG NOM/ACC-N
      "juopua" V NEG-PAST SG123
      "juopua" V PART-PERF SG123
```

The form *juopunut* is basically a participial perfect form of the verb *juopua*, but it is also the negative past form in all three singular persons. In addition, it is also an adjective in nominative or nominative accusative.

It seems that the adjective forms of the verb could be derived directly from the verb. However, there is a problem, because adjectives derived from verbs should be treated as adjectives with nominal inflection. Now such adjectives have double identification, as shown in (6).

(6)
```
"<juopuneena>"
      "juopua" V PART-PERF SG123 A SG ESS
"<juopuva>"
      "juopua" V PART-PRES A SG NOM
```

These forms can be corrected in post-processing, so that only the nominal inflection is left (7).

(7)
```
"<juopuneena>"
      "juopua" A SG ESS
"<juopuva>"
      "juopua" A SG NOM
```

With this method we can reduce the lexicon with thousands of lines, because we do not need to include such adjectives separately into the lexicon.

This method for describing adjectives applies to four types of adjectives, all derived from verbs. They are of the following type (8):

(8)
kävelevä 'walking'
kävellyt 'walked'
käveltävä 'walkable, to be walked, worth walking'
kävelty 'which has been walked'

The analysis gives the result as in (9).

(9)
```
"<kävelevä>"
      "kävellä" V PART-PRES
      "kävellä" V PART-PRES A SG NOM/ACC-N
"<kävellyt>"
      "kävellä" V NEG-PAST-SG
      "kävellä" V PART-PERF SG123
      "kävellä" V PART-PERF SG123 A SG NOM/ACC-N
"<käveltävä>"
      "kävellä" V PASS-PART-PRES
      "kävellä" V PASS-PART-PRES A SG NOM/ACC-N
"<kävelty>"
      "kävellä" V PASS/PASS-NEG-PAST
      "kävellä" V PASS A SG NOM/ACC-N
```

The last reading of each word is first analysed as a verb (V), and then analysis continues to nominal analysis, giving the analysis of an adjective (A). The adjective form of each word has ambiguous interpretation.
First we remove the verb reading (10).

(10)
```
"<kävelevä>"
      "kävellä" V PART-PRES
      "kävellä" A SG NOM/ACC-N
"<kävellyt>"
      "kävellä" V NEG-PAST-SG
      "kävellä" V PART-PERF SG123
      "kävellä" A SG NOM/ACC-N
"<käveltävä>"
      "kävellä" V PASS-PART-PRES
      "kävellä" A SG NOM/ACC-N
"<kävelty>"
      "kävellä" V PASS/PASS-NEG-PAST
      "kävellä" A SG NOM/ACC-N
```

Then we make separate readings for underspecified readings (11).

(11)

```
"<kävelevä>"
      "kävellä" V PART-PRES
      "kävellä" A SG NOM
      "kävellä" A SG ACC-N
"<kävellyt>"
      "kävellä" V NEG-PAST-SG
      "kävellä" V PART-PERF SG1
      "kävellä" V PART-PERF SG2
      "kävellä" V PART-PERF SG3
      "kävellä" A SG NOM
      "kävellä" A SG ACC-N
"<käveltävä>"
      "kävellä" V PASS-PART-PRES
      "kävellä" A SG NOM
      "kävellä" A SG ACC-N
"<kävelty>"
      "kävellä" V PASS
      "kävellä" V PASS-NEG-PAST
      "kävellä" A SG NOM
      "kävellä" A SG ACC-N
```

Now the analysis result of each word is ready for disambiguation.

There are also cases, where participial perfect form is in the function of noun. For example, *oppinut* (learned person) is often a noun. Such interpretations could be implemented as described above. However, the cases are so few that perhaps it is more economical to list them separately as nouns.

There is a minor problem in the above method for descibing adjectives derived from verbs. The lemma is the lemma of the verb ("*kävellä*"), and not of an adjective. It would be quite difficult, although not impossible, to control the adjectival lemma form in this complex process.

4 Describing compound nouns

The most common method of forming noun compounds in Finnish is to join nouns together, so that a noun compound forms a single word. This is not the only method, however, but the other methods do not concern us here. Single-word compounding is so productive that it is not possible to describe all compounds in the lexicon.

It would be tempting to allow free compounding of nous, but this would not work, because the system would jam into a never-ending loop.

The solution is somewhere in between. For example, it is possible to construct a separate lexicon and put there such nouns, which are likely to appear as first members in the compound. From this lexicon there would be access to the noun lexicon, and from the noun lexicon to inflection lexicons. This method works, because only the last member inflects. There are cases, however, where the first member is in genitive, but such words can be listed in the lexicon directly in genitive form.

This method works for most compounds. But we have a problem, if the compound has more than two members. Where is the boundary, if the word with more than two members should be cut into two parts? Consider the examples in (12).

(12)
a. *ulko-asiain-valio-kunta*
b. *valta-kunnan-kanslia*
c. *valtio-petos-rikos-oikeus*
d. *maan-puolustus-komitea*
e. *ihmis-oikeus-rikkomus*
f. *apulais-oikeus-asia-mies*
g. *perustus-laki-valio-kunta*
h. *perustus-laki*

We make a test and remove the first member to see whether the remaining word can appear as such. We see that (a) cannot, and also (b) and (c) are questionable. In the remaining cases it is possible.

Taking these considerations into account, we can split the words as in (13).

(13)
a. *ulkoasiain-valiokunta*
b. *valtakunnan-kanslia*
c. *valtiopetos-rikosoikeus*
d. *maanpuolustus-komitea*
e. *ihmisoikeus-rikkomus*
f. *apulais-oikeusasiamies*
g. *perustuslaki-valiokunta*
h. *perustus-laki*

Now in all cases the right side blocks are real words and can be listed in the noun lexicon, although some of them are compounds. The left side blocks are also compounds except (f) and (h).

This method of handling noun compounds condenses the lexicon a lot. There is a danger that the system would recognize as valid words also such strings that are not correct language. Tests show, however, that the danger is marginal.

5 Summary

I have shown that it is possible to construct a morphological analyzer of Finnish as a finite state transducer, without any morphophonological rules. The method requires that lexemes are classified into two groups according to their inflection type. Most words follow back vowel concordance and it is defined as default. Words with front vowel concordance are marked, and their inflection is direcred to front vowel inflection lexicons.

The omission of rules makes the lexicon complex, but on the other hand, there is no need to control the proper application of rules. The lexicon structure is transparent and easy to read.

The use of underspecification simplifies the lexicon structure somewhat, and underspecified readings can be 'read out' in post-processing phase to make disambiguation possible.

Particularly space-saving is the method of deriving many adjectives from verb stems. Using this method, there is no need to list such adjectives separately into the lexicon of adjectives.

The implementation of noun compounds uses an additional lexicon for such nouns, which are likely to be first members in noun compounds. These nouns are often single-word nouns, but also double-word nouns must be used in case of long compounds.